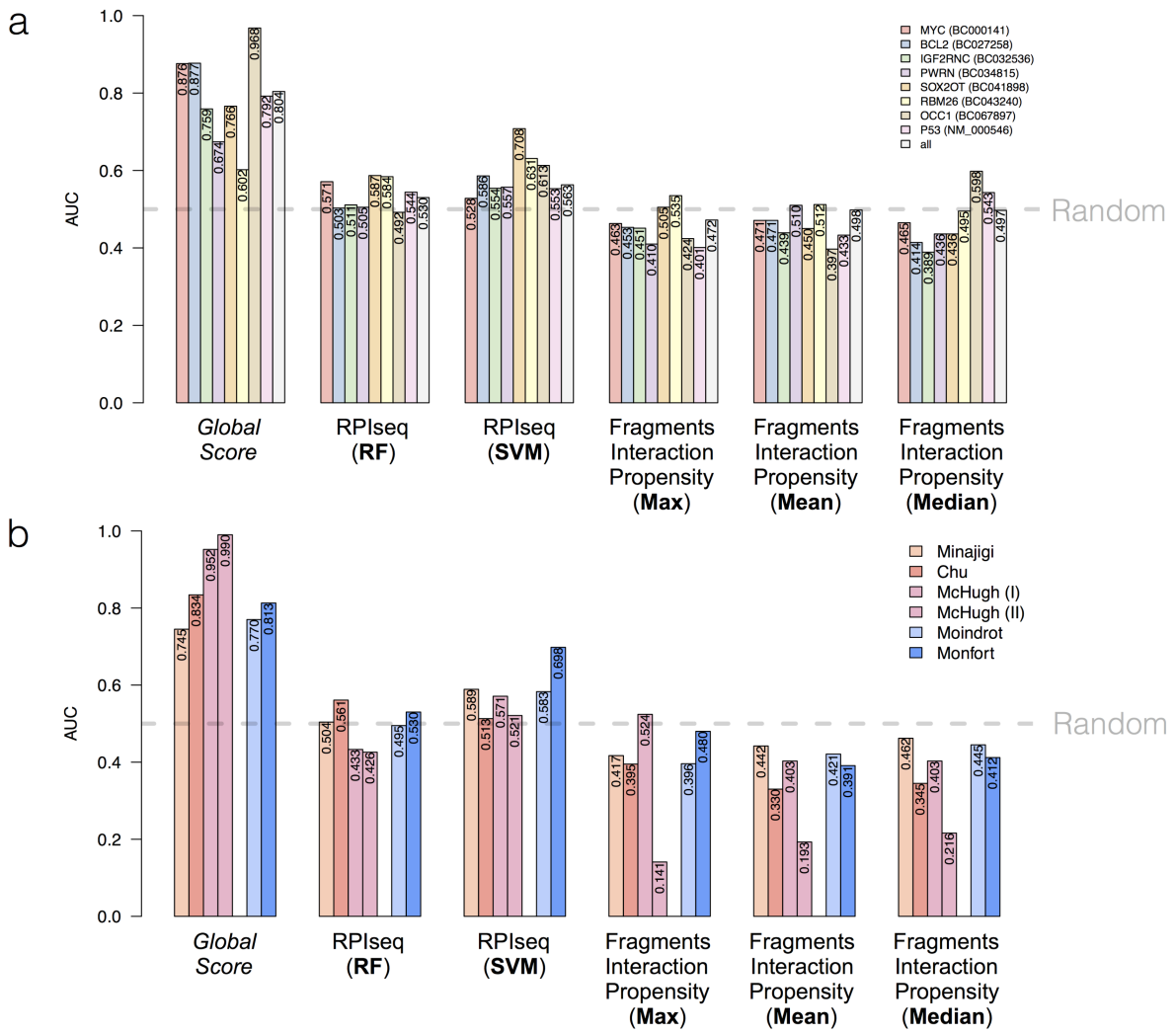
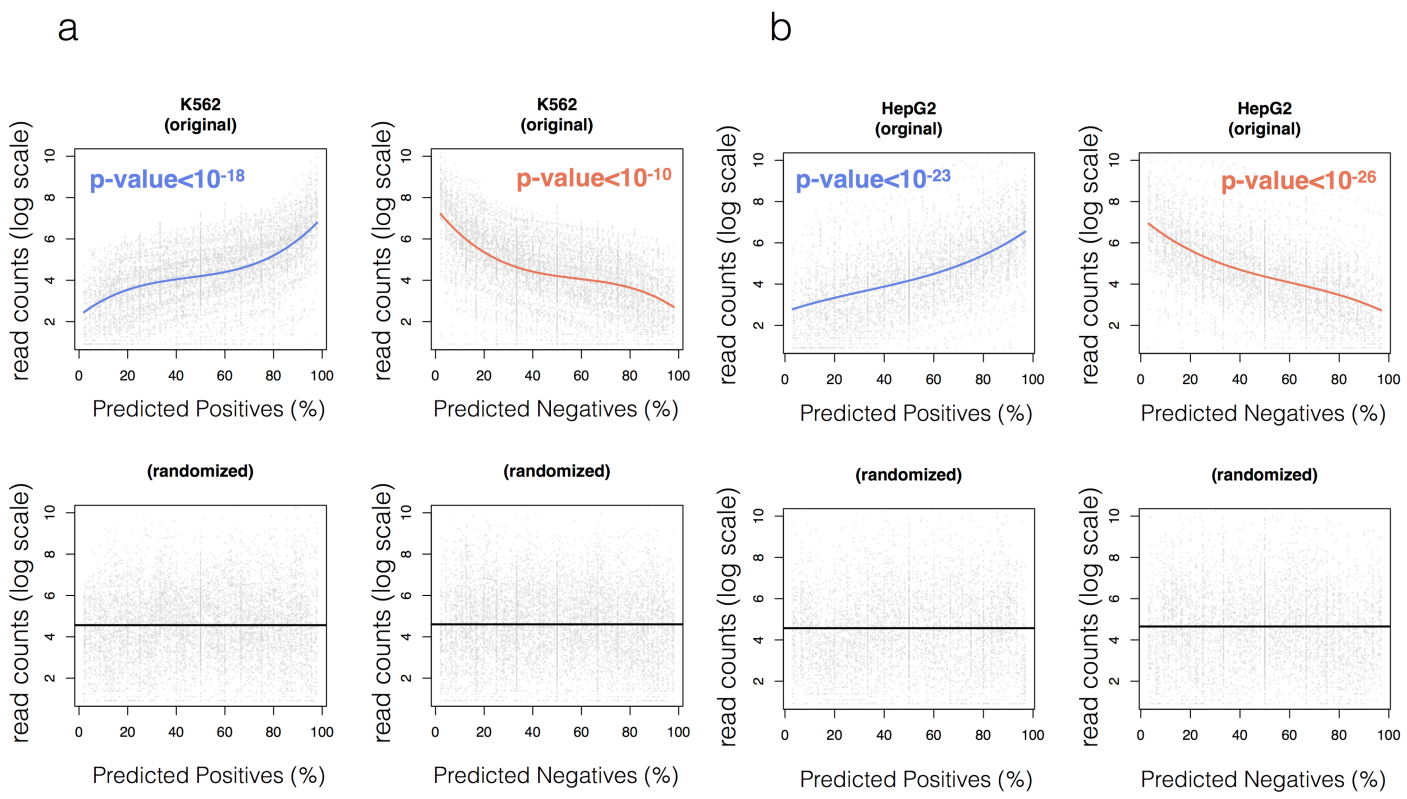


Supplementary Figure 1

Global Scores performances on the test sets and comparison with other methods

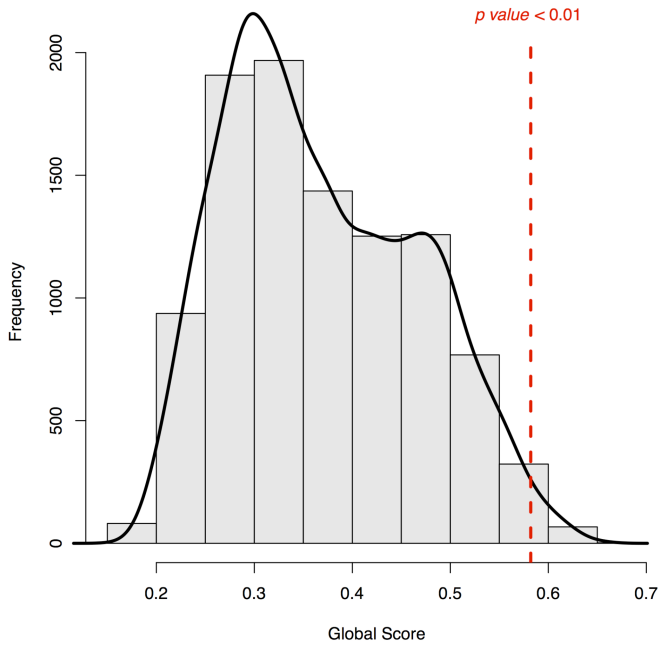
A) Performances of computational methods (Area Under the ROC curve AUC) on proteins-RNAs interactions revealed by protein microarray technology. B) *Xist* interactions with RBPs reported by Minajigi *et al.*, McHugh *et al.*, Chu *et al.* (proteomic studies) as well as Moindrot *et al.* and Monfort *et al.* (genomic studies). For each set of protein and RNA fragments, we measured mean, median and maximum of the interaction propensities calculated with *cat*RAPID and the binding score of *RPIseq*. *Global Score* outperforms *cat*RAPID-based analyses and RPIseq for large lncRNAs (more details in **Supplementary Tables 1-4**).



Supplementary Figure 2

Comparison between predicted and eCLIP-validated interactions.

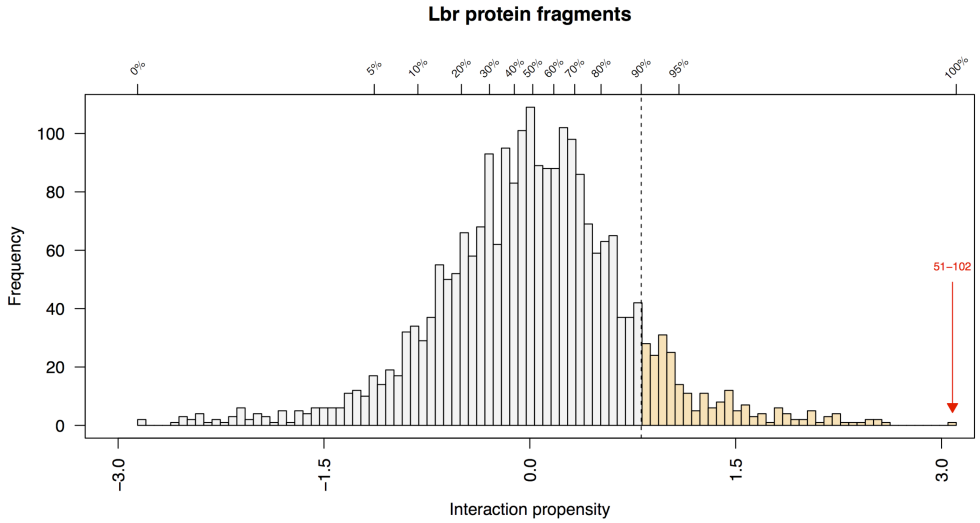
For 284 large transcripts (length >1000 nt), we studied the relationship between *Global Score* predictions and observed interactions revealed by eCLIP experiments in a) K562 and b) HepG2 cell lines. From low to high read counts, the fraction of interaction-prone RBPs (*Global Score* > 0.5) increases (upper plots; blue line) while RBPs with poor binding propensities (upper plots; red line; *Global Score* ≤ 0.5) show the opposite trend (log base 10 used for read counts; cubic function used for fitting). We assessed the significance of the trends by shuffling the read counts (bottom plots; black lines) and calculating two-sided Wilcoxon signed-rank test on predicted and randomized distributions. *Global Score* values are reported in **Supplementary Table 5**.



Supplementary Figure 3

Xist candidates selection.

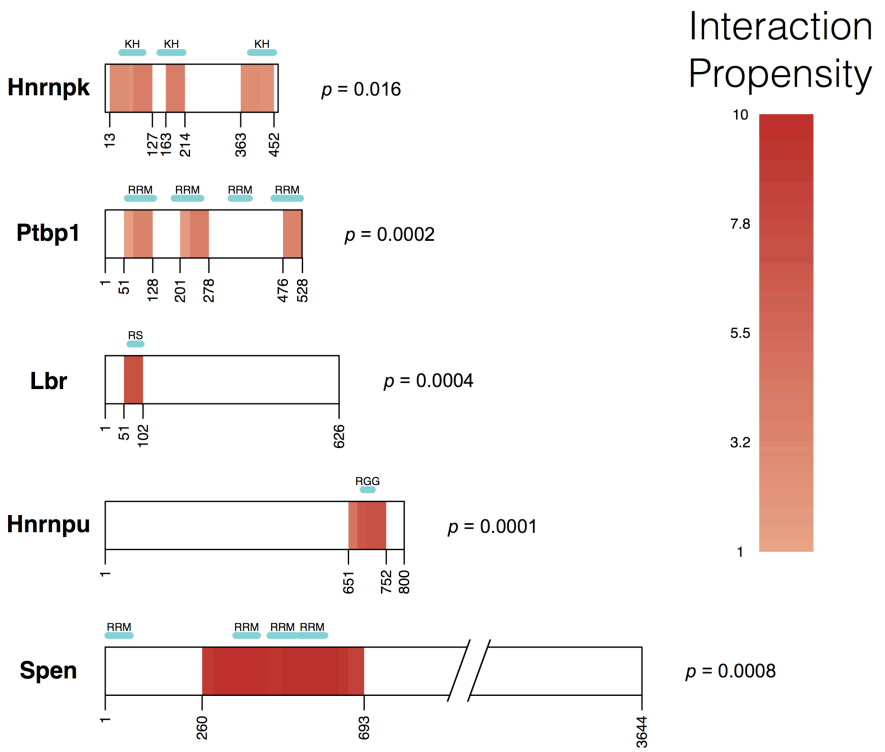
We randomized the association between *Global Score* values and number of independent experiments reporting *Xist* interaction with a specific RBP (> 600 proteins used for the analysis; 10000 randomizations performed; see also **Fig. 1d**). *Global Score* values above 0.59 significantly discriminate 38 RBPs reported in at least two experimental assays (empirical p-value<0.01; **Supplementary Table 6** and **7**).



Supplementary Figure 4

Predictions of the RNA-binding domain of Lbr.

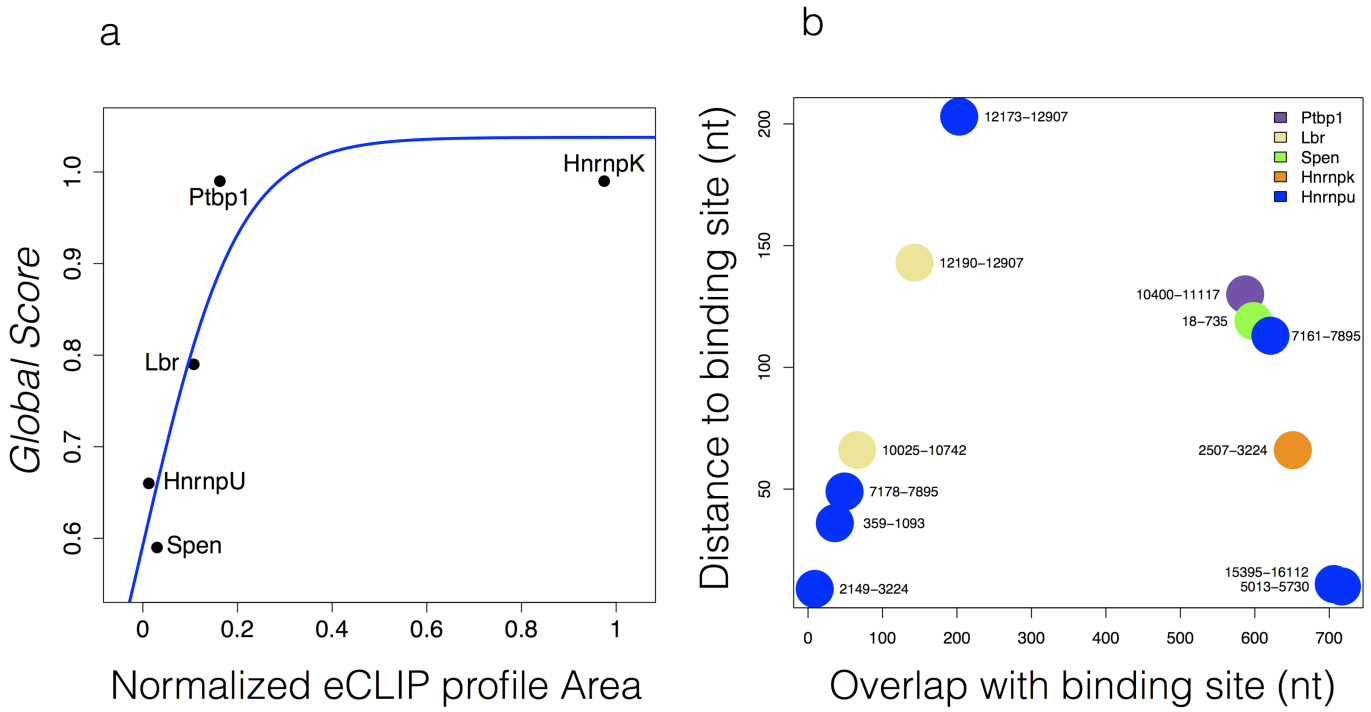
We ranked Lbr fragments by their interaction propensity to *Xist* lncRNA. The fragments corresponding to the top 10% of the statistical distribution are highlighted in yellow. The highest interaction propensity corresponds to amino acids 51-102 which corresponds to the RS domain implicated in nucleic acid recognition.



Supplementary Figure 5

RNA-binding regions of Spen, Hnrnpk, Hrnnpu/Saf-A, Lbr and Ptbp1.

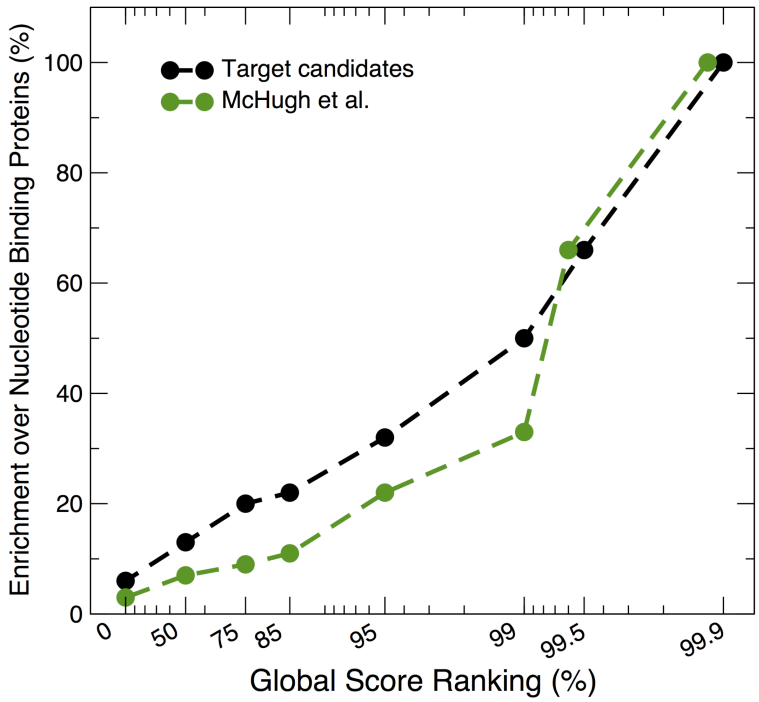
Fragments overlapping with RNA-binding domains (RRM, KH, RGG and RS) rank high (top 2%) with respect to other protein regions (empirical p-values reported on the right). Fragments with the highest scores (top 2%) are coloured according to their interaction propensities.



Supplementary Figure 6

Predicted vs validated binding sites.

A) Relationship between *Global Score* values and areas under eCLIP profiles (Pearson correlation of 0.93 using the fitting formula *Global Score* = *α* tanh (eCLIP) + *β;* p-value = 0.02). The areas are normalized relatively to the largest value of HnrnpK. B) Proximity of predicted binding sites to eCLIP peaks evaluated in terms of distance and overlap. Predicted fragments are in close proximity of eCLIP peaks and overlapping with them (significance of predictions is reported in **Fig. 1e)**. The maximum distance observed (200 nt) is below the average distance between overlapping fragments (367 nt) and the maximum overlap corresponds to the fragment size (718 nt) used in our analysis.



Supplementary Figure 7

Significance of Global Score predictions.

We compared interaction propensities of target candidates with a large set of nucleotide-binding proteins. From low to high *Global Score* values, the ratio of identified candidates over number of predicted interactions increases monotonically, reaching 50% at the 99th *Global Score* percentile (p-value = 10^{-8}) and 100% at the 99.9th percentile (p-value = 10^{-20}; **Supplementary Table 9**).

Supplementary Methods

Local predictions of protein-RNA interactions

We previously developed *cat*RAPID to predict the interaction propensity of protein and RNA sequences using their physico-chemical properties[1,2]. The method, which was designed to complement experimental studies, has an average accuracy of 78% in predicting binding partners and works for transcripts shorter than 1000 nt due to the difficulty of modeling the structure of larger sequences. Indeed, the size of the configuration space makes structural predictions difficult for thermodynamic approaches.

Previous pilot projects indicate that division of sequences into sub-elements is useful to identify contacting regions (section *Binding sites predictions*). For instance, by fragmenting protein and RNA sequences, it is possible to detect the binding sites of Fragile X mental retardation protein FMRP and TAR-DNA binding protein 43 TDP-43[3]. Yet, when proteins bind with low affinity to multiple regions of RNA sequences, identification of binding regions cannot be directly exploited to predict the binding strength between two molecules. For instance, Histone-lysine N-methyltransferase Ezh2 is predicted to associate with *Xist* in several sites within the repetitive region A, but the interactions have low interaction propensities (section *catRAPID predictions of Polycomb Repressive complex proteins PRC2 interactions*)[4].

For each protein and RNA fragment, contributions of secondary structure, hydrogen bonding and van der Waals' are combined into the *interaction profile*[1]:

$$\vec{\Phi}_x = \alpha_H \vec{H}_x + \alpha_W \vec{W}_x + \alpha_S \vec{S}_x \qquad (1)$$

where the variable x indicates RNA ($x = r$) or protein ($x = p$). The hydrogen bonding profile, denoted by $\vec{H}$, is the hydrogen bonding ability of each amino acid (or nucleotide) in a protein (or RNA) sequence:

$$\vec{H} = H_1, H_2, \ldots, H_{length} \qquad (2)$$

Similarly, $\vec{S}$ represents the secondary structure occupancy profile and $\vec{W}$ the van der Waals' profile. The *interaction propensity* π is defined as the product between the protein propensity profile $\vec{\Psi}_p$ (Fourier's transform of $\vec{\Phi}_p$) and the RNA propensity profile $\vec{\Psi}_r$ (Fourier's transform of $\vec{\Phi}_r$)

weighted by the *interaction matrix* I (coefficients are provided in our previous publication[1]):

$$\pi = \vec{\Psi}_p \mathrm{I}\, \vec{\Psi}_r \quad (3)$$

In our approach, polypeptide and nucleotide sequences are divided into overlapping fragments followed by prediction of individual interaction propensities.

Global Score

A key problem with prediction of global features of polypeptide and nucleotide chains is the integration of the signal derived from local properties. While knowledge of features encoded by fragments is informative, the overall context should be taken into account to accurately predict interaction abilities.

We implemented a non-linear algorithm that integrates the information contained in the interaction propensities of protein and RNA fragments. To train the method we used different sets of binding (positives) and non-binding (negatives) protein-RNA pairs. The classification into positives and negatives allows us to make predictions independently of the statistical distributions of experimental scores that are intrinsically linked to each individual technique, thus ensuring wide applicability of the approach.

We trained *Global Score* on PAR-/HITS-CLIP interactions of Ago1, Ago2, Ago4, Elavl1, Qki, Pum1, Pum2, Tnrc6a, Tnrc6b, Tnrc6c, Ncl, Igf2bp1, Igf2bp2 and Igf2bp3, which were measured in similar experimental conditions and are annotated in AURA (UTR lengths > 1000 nt)[5]. To avoid biases toward cases with larger number of partners, we selected a fixed number of sequences (50 RNAs) for each protein in the positive set. We shuffled RNA interactions of the RBPs and selected the same number of cases to build a balanced negative set (50 RNAs per protein; **Supplementary Tables 1** and **2**). In our analysis, we filtered out similar RNA sequences using CD-HIT (http://weizhongli-lab.org/cd-hit/; sequence identity > 80%; **Supplementary Tables 1** and **2**).

Once protein and RNA sequences are generated with the fragmentation procedures[3,4], the distribution of the interaction propensity scores π (Eq. 3) is computed:

$$f_i = \vartheta(\pi - i)[1 - \vartheta(\pi - i - 1)] \quad (4)$$

where $\vartheta(x)$ is the Heaviside function that is 1 if $x > 0$ and zero otherwise. The values f_i are weighted to norm 1:

$$F_i = f_i / \sum_{i=min}^{max} f_i \quad (5)$$

where $min = -50$ and $max = 50$. To determine the relative contribution F_i of fragments, we computed h_k:

$$h_k = tanh\left(\omega_k^i F_i\right) \quad (6)$$

where $tanh(x)$ is the hyperbolic tangent of x. The global score Π is evaluated using h_k:

$$\Pi = tanh\left(\Omega^k h_k\right) \quad (7)$$

The weights ω_k^i and Ω^k have been determined by optimizing the match between experimental and predicted interactions. To avoid over-fitting, we varied the number of internal weights proportionally to the size of the training set and performed a 5-fold cross-validation at each optimization. On a 5-fold cross-validation, we obtained an AUC of 0.84 (**Fig. 1b; Supplementary Fig. 1; Supplementary Tables 1** and **2**) in discriminating interacting and non-interacting protein-RNA pairs. We note that having a continuous range in the score of the algorithm ensures flexibility in the training phase, as the use of a binary score would increase the number of unclassifiable cases (in between the two states). The identification of a cut-off through the ROC analysis (Youden's index=0.5) provides the optimal score to discriminate interacting vs non-interacting protein-RNA pairs.

We performed an independent validation using 8 transcripts > 1000 nt (*Myc, Bcl2, Igf2rnc, Pwrn1, Sox2oy, lincRBM26, Occ1* and *Tp53*) whose binding partners have been determined through protein microarrays technology[6]. For each RNA molecule, we selected 50 top-ranked (i.e., high-affinity) and 50 bottom-ranked (i.e., low-affinity; **Supplementary Tables 1** and **2**) RBPs, carried out the fragmentation, as described in our previous publication[4], computing the overall interaction propensities with the *Global Score* method. We observed high performances on the protein array test set (AUC=0.80; **Fig. 1b**; **Supplementary Tables 1, 2** and **3**; section *Comparison with other methods*). The analysis includes 20 non-canonical RBPs (**Supplementary Table 4**)[7] that were correctly predicted to bind to their targets in 75% of the cases (15 out 20 RNAs), which suggests

that *Global Score* is not biased towards known RNA-binding domains. Detailed testing performances are reported in the section *Comparison with other methods.*

Large transcript analysis

We compared *Global Score* predictions of protein-RNA interactions with data from eCLIP assays (161 experiments including replicas; total 60 RBPs downloaded in February 2016 of which 32 studied in HepG2 cell line and 48 in K562 cell line; BED files from https://www.encodeproject.org)[8]. For each protein, we ranked the target genes by number of reads and selected 5 transcripts >1000nt with the largest amount of total counts (284 RNAs). We subsequently collected the read counts of each transcript in all the eCLIP assays.

We applied *Global Score* to all RBP-RNA pairs (60x284 predictions) and studied the relationship between our calculations and the read counts in HepG2 and K562 cell lines. We observed that the number of predicted interactions significantly increases with the read counts (**Supplementary Figure 2**), while pairs that are predicted to not interact show the opposite trend. To quantify the statistical significance of the results, we shuffled the read counts within the pool of proteins associated with each transcript (K562 cell line; p-value < 10^{-47} for predicted positives and p-value < 10^{-28} for predicted negatives; HePG2 cell line: p-value < 10^{-17} for predicted positives and p-value < 10^{-4} for predicted negatives; two-sided Wilcoxon's signed-ranked test). Our analysis indicates that there is a significant relationship between our calculations and the binding strengths.

Xist database generation

Recent publications created an unprecedented wealth of information on *Xist* interactions as well as functional players in XCI[9-13]. Minajigi *et al.*[11] and McHugh *et al.*[9] exploited oligos complementary to *Xist* to recover interacting partners using UV-crosslinking conditions (iDRiP, RAP-MS). They found, respectively, about 250 and 20 direct *Xist* interactors. Similarly, Chu *et al.*[10] used formaldehyde crosslinking and mass-spectrometry to identify 81 *Xist* interactors (ChIRP-MS). Minajigi *et al.*[11] and McHugh *et al.*[9] identified *bona fide Xist*-interactors using a zero-length crosslinker agent (UV-crosslinking) and denaturing conditions for the biochemical purification of *Xist*-interacting partners. Chu *et al.*[10] revealed direct and possible indirect associations as the experimental protocol employed formaldehyde-crosslinking, which fixes associations within ~2 Å

radius and they used non-denaturing purification conditions which may allow non-direct *Xist*-interacting proteins to be recovered.

Moindrot *et al.*[12] and Monfort *et al.*[13] used loss-of-function genetic screens to extrapolate functional *Xist* silencing partners. Moindrot *et al.*[12] developed a conventional shRNA screen, using an inducible *Xist* Embryonic Stem cell (ESC) reporter line, while Monfort *et al.*[13] took advantage of insertional mutagenesis screen in a previously established haploid ESC. Moindrot *et al.*[12] focused on cells in which *Xist* fails to silence an *in cis* GFP-reporter gene upon shRNA transduction. Monfort *et al.*[13] relied on cell survival to measure *Xist* inability to trigger silencing of the only X chromosome upon viral gene-trap insertion.

While proteomic approaches[9-11] can be exploited to reveal proteins binding to *Xist*, they do not differentiate between functional *Xist*-interactors and other house-keeping functions (RNA-processing, polyadenylation, etc). By contrast, loss-of-function genetic screens select important regulators of XCI, but fail to provide information of direct protein interactions[12,13]. Moreover, due to their experimental set-up genetic screens are devoid of proteins that interfere with cell proliferation or cell survival[14]. We also reanalyzed the raw data from the genetic screening by Moindrot *et al.*[12] . The effect of shRNAs targeting specific genes was calculated by dividing final counts ("sorted") over initial counts ("input"). The ratio of each individual shRNA was standardized by subtracting the median ratio of the dataset followed by division with median absolute deviation. The third highest standardized ratio of shRNAs targeting the same gene was used as score for the ranking. At least three individual shRNAs show higher or equal enrichment in counts were employed to assure consistent results and avoid off-targets shRNAs. The overlap between Moindroit *et al.*[12] with proteomic datasets (342 genes combining data from Chu *et al.*[10], McHugh *et al.*[9] and Minajigi *et al.*[11]) is 17 genes. Ranking by the third highest standardized ratio (top 300 genes) we identified 18 genes (Cdkn2a, Hnrnpl, Khsrp, Lox, Matr3, Mcm3, Msh2, Numa1, Nxf1, Pcbp2, Ptbp1, Rbm15, Sap18, Spen, Thoc2, Trp53, Wdr33, Wtap). The overlap between the 22 genes listed by Monfort *et al.*[13] and the proteomic datasets is of 1 gene (Spen).

To summarize, in our analysis we used: the top 300 genes from Moindrot *et al.*[12] and 22 genes (21 proteins and 1 noncoding RNA) from Monfort *et al.*[13]. Proteomic screens comprise 81 genes (81 proteins) from Chu *et al.*[10], 1768 genes (1767 proteins and Q6ZWY8 < 50 aa; 300 high-confidence hits) from Minajigi *et al.*[11] and 20 genes (20 proteins) from McHugh *et al.*[9] (**Supplementary Table 6**; section *Xist candidates selection*). We considered as negatives all the genes that Minagiji *et al.*

found enriched in male vs. female cells [29 proteins with female vs male fold change $\log_2(FC) < -1.0$][11]. As for the datasets by Chu *et al.*[10]. McHugh *et al.*[9] and Monfort *et al.*[13], we retrieved protein sequences from Uniprot using gene names (http://www.ebi.ac.uk/reference_proteomes). In the case of Moindrot *et al.*[12] and Minajigi *et al.*[11], we employed the protein identifiers (Moindrot RefSeq IDs were converted to UniProt IDs with 100% sequence similarity).

Comparison with other methods

Global score is the first algorithm to quantitatively predict RBP partners of RNA>1000nt. Indeed, *cat*RAPID and *lncPro*[15] cannot be applied to predict protein interactions with large transcripts because the secondary structure of the RNA is calculated with thermodynamic-based approaches (sequence size < 1000nt)[16]. Another method, *RPIseq*[17], computes protein-RNA interactions based on amino acid and nucleotide frequencies (two classifiers are available: Random Forest, RF, and Support Vector Machine, SVM). On the test set, *RPIseq* shows lower performances (*RPIseq* RF/SVM: AUC of 0.53/0.56, specificity of 0.31/0.43, sensitivity of 0.74/0.68 and MCC of 0.12/0.06; **Supplementary Table 3**) than *Global Score* (AUC of 0.80, specificity of 0.71, sensitivity of 0.78 and MCC of 0.44; **Supplementary Table 3**), which suggests that sequence patterns do not capture the physico-chemical determinants of binding. To assess to what extent the use of a non-linear algorithm is effective for the integration of the signal coming from protein and RNA fragments, we measured mean, median and maximum of the interaction propensities computed as defined in Eq. 3. On the test set, we observed lower performances (mean/median/max of interaction propensities: AUC of 0.49/0.49/0.47, specificity of 0.21/0.18/0.44, sensitivity of 0.80/0.83/0.63 and MCC of 0.02/0.02/0.07), indicating that *Global Score* is more efficient than methods based on the simple statistical analysis of interaction propensities.

In summary, *RPIseq* and fragments statistics show a preference for predicting positive interactions (sensitivities > 0.7), but fail to recognize negatives (specificities <0.5). Similar results were observed for *Xist* interactions: *Global Score* (considering all the experiments: AUC of 0.77, specificity of 0.96, sensitivity of 0.55 and MCC of 0.41; **Supplementary Table 3**) outperforms the other approaches (*RPIseq* RF/SVM: AUC of 0.50/0.58, specificity of 0.20/0.65, sensitivity of 0.82/0.53 and MCC of 0.03/0.15; mean/median/max: AUC of 0.48/0.48/0.40, specificity of 0.24/0.63/0.72, sensitivity of 0.85/0.44/0.55 and MCC of 0.09/0.06/0.22), although it must be noted that specific datasets are associated with different performances (**Supplementary Table 3**).

In agreement with experimental evidence, 97% (28/29) of the genes that Minagiji *et al*. found enriched in male vs female cells ($\log_2(FC) < -1.0$)[11] are predicted by *Global Score* as non-interacting, while *RPIseq* is not able to identify them (RF: 0/29; SVM: 0/29). By expanding the list of negative candidates (female vs. male fold change $\log_2(FC) < -0.5$), *Global Score* correctly identifies 81% of the proteins (173/214), while *RPIseq* shows specificity closed to zero (RF: 0/214; SVM: 1/214).

Xist candidates selection

We sought to determine which of the proteomic and genetic candidates (623 proteins) are direct *Xist* interactions. *Global Score* calculations indicate that the two datasets by McHugh *et al.*[9] (published I and unpublished results II) are associated with the highest predictive power (Area under the ROC curve AUC of 0.95 for I and 0.99 for II; **Fig. 1c; Supplementary Table 3**) followed by Chu *et al.*[10] (AUC=0.83), Monfort *et al.*[13] (AUC=0.81), Moindrot et al.[12] (AUC=0.77) and Minajigi *et al.*[11] (AUC=0.74).

We observed that *Global Score*, which ranges between 0 and 1, significantly correlates with the number of experiments reporting interaction of a specific gene with *Xist* (i.e. hits found in multiple studies have higher values; **Fig. 1d**). This finding indicates that there is a tight link between the experimental reproducibility and the computational evidence of an interaction. Upon randomization of the number of experiments associated with a specific hit, the *Global Score* threshold of 0.59 significantly differentiates genes reported in at least two experiments and the rest of associations (empirical p-value<0.01 calculated on 10^4 randomizations; **Supplementary Fig. 3**).

Selecting genes appearing above the threshold of 0.59 (58 candidates; **Supplementary Table 2**) and reported in at least two independent datasets, we identified 38 proteins with medium- (i.e., $0.59 \leq$ *Global Score* ≤ 0.80) and high- ($0.80 <$ *Global Score* ≤ 1, marked) interaction propensities (**Supplementary Tables 6** and **7**). None of the target candidates is predicted to interact with U1 RNA (**Supplementary Table 7**), in agreement with the RAP-MS experiments reported by McHugh *et al.* [9] (by contrast Snrpd2 and Snrpe are correctly predicted as U1 interactors). Notably, 29 out of 38 proteins have high interaction propensities and 20 are associated with *Global Score* ≥ 0.95.

We screened our candidates for cellular localization (i.e. direct interactions in the nucleus), functional categories (i.e. RNA metabolism, gene-silencing), protein association networks (i.e.

STRING interactions; section *Interaction network*) and expression-levels (i.e. expressed in early embryogenesis, **Supplementary Tables 6** and **7**).

GO analysis reveals that 21 out of the 38 candidates are part of the Hnrnp protein network (**Supplementary Table 7**). HnrnpU and HnrnpK are key regulators of X-Chromosome inactivation: they are necessary for *Xist*-localization to chromatin (and, in turn, gene-silencing) [9,18] and Polycomb recruitment, respectively[10]. We also found a sub-network including Rbm15, Spen and Rbm3 that is involved in Ncor-complex recruitment to the inactive X[9,10].

Almost all of the 38 candidates are in the RNA-related functional categories (35 out of 38 genes). We observed functional associations with RNA-related processes, especially post-transcriptional regulation, splicing and nuclear trafficking. The last category is particularly interesting as *Xist* is a poly-adenylated spliced RNA that never leaves the nucleus[19]. More than half of the selected genes (20 out of 38; **Supplementary Table 7**) are associated with the transcriptional regulation category. Other candidates are part of the silencing machinery (Ncor2 / Spen and Hdac1 complex / Rbm14) or are important for RNA processing and stabilization (Hnrnp-proteins). Three out of 38 genes are also part of the nuclear matrix (Lbr, Matr3, HnrnpM), a sub-compartment that contacts *Xist* and is involved in gene silencing.

To infer functional relationships among the selected candidates, we clustered the initial pool of 58 genes based on enriched GO terms of direct interactions (section *Gene ontology clustering*). We identified two major groups: one related to RNA splicing and transport, and another related to transcription regulation and protein degradation. The two classes contain genes that are important for *Xist* spreading and localization to the chromatin (Hnrnpu/Saf-A)[19] and are relevant for *Xist* localization to the nuclear lamina (Lbr) and may be relevant for *Xist* localization to the nucleolus[20].

Interaction network

The network of protein-protein interactions was built using STRING (http://string-db.org/), selecting confidence scores ranging between 0.70 and 0.90. Most of the interactions are reported with confidence score of 0.90. Interactions among Spen, Rbm15 and Rbm3 have been manually curated (**Supplementary Table 7**)[9,10].

Gene ontology clustering

We clustered candidate genes using functional macro-categories of interest ("Chromatin remodeling", "Nuclear matrix and envelop", "RNA processing and splicing", "Transcription regulation"). Gene Ontologies (GO) terms are assigned to a macro-category querying their definitions using keywords (i.e. the words in the macro-category; **Supplementary Table 7**).

catRAPID predictions of Polycomb Repressive complex proteins PRC2 interactions

Polycomb Repressive complex proteins PRC2 did not appear in our analysis. This is due to the fact that PRC2 elements were not over-represented in proteomic[9-11] or genetic screens[12,13]. In agreement with these findings, we run calculations for PRC2 elements and observed low interaction propensities (Suz12: *Global Score* = 0.01; Ezh2: *Global Score* = 0.35).

In our previous publication[3], the *cat*RAPID approach was used to assess the interaction ability of PRC2 components to *Xist* regions (overall interaction propensity was not possible as outlined in *Local predictions of protein-RNA interactions*). Using randomized RepA as a control, Ezh2 was predicted to bind *Xist* with medium specificity (interaction strength = 75%) and low affinity (Ezh2-RepA interaction propensity < 1). These findings are in good agreement with recent 3D-SIM data, showing poor overlap between *Xist* and PRC2[21]. In fact, recent data by STochastic Optical Reconstruction Microscopy (STORM)[22] indicate non-random association of *Xist* and PRC2. *Xist* and PRC2 are closer than expected by chance, but the interaction is unstable and therefore may not be biologically relevant or mediated by other proteins[21].

Binding sites predictions

As shown in our previous studies, the fragmentation procedure identifies RNA-binding regions in detail[3,4] . In the case of Lbr, fragmentation identifies amino acids 51-102 as the most prone to interact with RNA, which is in agreement with the annotation of the RS domain involved in nucleic acid recognition[23] (**Supplementary Fig. 4**). To predict regions contacted by RBPs, we use fragments whose annotation is compatible with the RNA-binding domains (RBDs) reported in Gerstberger *et al.*[24] and NPIDB ('RNA' and 'hybrid' families; update September 2015; http://npidb.belozersky.msu.ru/). The fragments were ranked to identify high-confidence regions (**Supplementary Fig. 5**). For the five proteins used in this study, RBD-containing fragments are

predicted to be the most interaction prone (top RDB-containing fragments rank in the range 1% to 0.0001%; **Supplementary Fig. 5**).

We introduced the *signal localization* procedure to reveal significant interactions among those generated while fragmenting protein and RNA sequences (average of 4000 interactions between *Xist* and each of the eCLIP candidates). The interactions are computed with uniform fragmentation[3], which samples all regions within *Xist*. For each protein, we selected the highest-scoring interactions (top 2%). We then computed the midpoint and distance of each selected fragment from the midpoint. We discarded fragments when distance is > 5 times the size of the RNA fragment, which is the resolution of the method. The fragments are reported in Fig. 1e and their significance is estimated by randomization (sensitivity of the predicted hits with respect to 10^4 random predictions / protein). To distinguish between localized and dispersed signals, we introduced the concept of *signal dispersion* that is defined as 2 times the standard deviation of the distance between fragments. If the *signal dispersion* is larger than the resolution, all the highest-scoring fragments are considered for the statistical analysis. Hnrnpu/Saf-A shows the largest *signal dispersion* (2500 nt, while Spen, Lbrm HnrnpK and Ptbp1 have respectively: 1200, 1500, 1700 and 1800 nt), which indicates that binding is less specific, as also revealed by eCLIP experiments (**Fig. 1e**).

Prediction and validation of Xist interactions

We analysed representative candidates with different *Global Score* values ranging from the minimal cut-off (0.59) to the highest propensity (0.99). HnrnpU and Spen have a role in *Xist*-mediated silencing and are associated with medium interaction propensities (Hnrnpu/Saf-A *Global Score* = 0.66 and Spen *Global Score* = 0.59); Lbr and HnrnpK have been described to have a role in gene silencing and Polycomb recruitment, respectively, and show higher *Global Score* values (Lbr *Global Score* = 0.79 and HnrnpK *Global Score* = 0.99). Ptbp1 has the highest interaction propensity (*Global Score* = 0.99), although its role in XCI establishment is not yet known[9,10]. Dkc1 is use as negative control (*Global Score* = 0.01) and is not known to be involved in XCI.

We found a tight correlation between *Global Score* values and eCLIP profile areas (Pearson correlations of 0.87 fitting with $Global\,Score = \alpha \log(\text{eCLIP}) + \beta$ and 0.93 using $Global\,Score = \alpha \tanh(\text{eCLIP}) + \beta$; p-value < 0.02 (t-test, t-value = 4.337, DF = 3); **Supplementary Fig. 6a**). Profiles with a high peak height and a small peak base (HnrnpK) or a moderate peak height and a large peak base (Ptbp1) have strong *Global Score* values (~0.9).

Profiles with a moderate peak height and a medium-size peak base (Lbr) have intermediate *Global Score* values (~0.7). Profiles with a low peak height and a large peak base (HnrnpU) or a moderate peak height and a small peak base (Spen) have weak *Global Score* values (0.59). The correlation between Global Score and the profile area is indicative of *Global Score* ability in providing a quantitative estimate of protein-RNA interaction affinity.

Predicted binding regions of Hnrnpk, Hnrnpu/Saf-A, Lbr, Ptbp1, and Spen have been validated with eCLIP (section *eCLIP experiments*). Highest scoring associations overlapping with annotated binding domains (>50% coverage) are: Hnrnpk (P61979) 363-414 aa (KH domain) interacting with Xist 2507-3224 nt (0.98 percentile, **Supplementary Fig. 5**); Hnrnpu/Saf-A (Q8VEK3) 700-751 aa (RGG domain) with Xist 376-1093 nt (0.99 percentile, **Supplementary Fig. 5**); Lbr (Q3U9G9) 51-102 aa (most interacting Lbr fragment, **Supplementary Fig. 4** and **5**) with 10025-10742 nt (0.98 percentile, **Supplementary Fig. 5**); Ptbp1 (P17225) 76-127 aa (RRM domain) with Xist 10741-11458 nt (0.99 percentile, **Supplementary Fig. 5**); Spen (Q62504) 332-477 aa (RRM domain) with Xist 18-735 (0.98 percentile, **Supplementary Fig. 5**).

We ranked the regions containing predicted binding sites using the signal-to-noise ratio (SNR) of interaction propensities (**Supplementary Table 8; Fig. 1e**). Fragments in regions with high SNR show either close proximity to experimental binding regions or overlap with them (**Supplementary Fig. 6b**). Indeed, the average distance between experimental and predicted binding regions is 224 nucleotides, which is consistent with the resolution of our method. For each protein, we assessed the significance of the match between predicted and validated binding sites by randomizing their association 10^4 times and measuring the true positive rate.

Global Score as a tool to prioritize candidate from functional annotations

In addition to predicting interactions in datasets enriched in physical[9-11] or functional associations[12,13], we used *Global Score* to identify binding partners in a pool of nuclear proteins. In this analysis, we employed a reference set of 532 nucleotide-binding proteins linked to the GO category "Nucleotide Binding" and annotated as "Nuclear" (GO:0000166). To evaluate the ability of *Global Score* to identify interactions *de novo*, we measured the ranking of our 38 target candidates as well as all the hits identified by McHugh *et al.*[9]. We use all the proteins reported by McHugh *et al.* as a control, as they are associated with high-confidence predictions (AUC > 0.90)[9-11].

From low to high *Global Score* values we measured the enrichments by calculating the ratio of identified targets over the number of predicted interactions. All the 38 target candidates were found above the 50^{th} percentile of the *Global Score* (p-value = 0.0045; Fisher's exact test) and the enrichments showed a monotonic increase reaching 50% at the 99^{th} *Global Score* percentile (p-value = 10^{-8}; Fisher's exact test) and 100% at the 99.9^{th} percentile (p-value = 10^{-20}; Fisher's exact test; **Supplementary Fig. 7** and **Supplementary Table 9**).

The two target candidates Rbm3 and HnrnpK (99.5^{th} percentile, p-value < 0.00025; Fisher's exact test) showed the highest *Global Score* values. Among the 10 proteins associated with the highest interaction propensities, we found Aurka, Rmb34 and Rmb38. Intriguingly, Aurkb has previously been found to regulate *Xist* retention on metaphase chromosomes during cell-cycle progression[25]. It is possible that Aurka, which forms a complex with Aurkb, plays a role in *Xist* spreading. Rbm34 was previously identified in the screening by Moindrot *et al.*[12] (position 646 out of top 1000 ranked genes).

We obtained similar performances on the datasets by McHugh *et al.*[9] (**Supplementary Fig. 7**) suggesting that *Global Score* can be used as a tool to enrich for RNA direct binders in large datasets. We note that the approach shows remarkable performances despite not taking into account the physiological abundance of proteins, which might prevent some of the physical interactions from occurring in the cellular environment.

eCLIP experiments

We crosslinked 6 hours doxycycline-induced pSM33 mouse male ES cells with 0.4J of UV254. Cells were lysed in 1 ml lysis buffer (50 mM Tris pH 7.5, 100mM NaCl, 1% NP-40, 0.5% Sodium Deoxycholate, 1x Protease inhibitor cocktail). RNA was digested with Ambion RNase I (1:4000 dilution) to achieve a size range of 100-500 nucleotides in length. Lysate preparations were precleared by mixing with Protein G beads for 1hr at 4C. Target proteins were immunoprecipitated from 5 million cells with 10 ug of antibody and 75 ul of Protein G beads in 100uL lysis buffer. The antibodies were pre-coupled to the beads for 1 hr at room temperature with mixing before incubating the precleared lysate to the beads-antibody overnight at 4C. After the immunoprecipitation, the beads were washed four times with High salt wash buffer (50 mM Tris-HCl pH 7.4, 1 M NaCl, 1 mM EDTA, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate) and four

times with Wash buffer (20 mM Tris-HCl pH 7.4, 10 mM MgCl2, 0.2% Tween-20). RNAs were then eluted by incubating at 50C in NLS elution buffer (20 mM Tris-HCl pH 7.5, 10 mM EDTA, 2% N-lauroylsacrosine, 2.5 mM TCEP) supplemented with 100 mM DTT for 20 minutes. Samples were then run through a standard SDS-PAGE gel and transferred to a nitrocellulose membrane, and a region 75 kDa above the molecular size of the protein of interest was isolated and treated with Proteinase K (NEB) followed by buffer exchange and concentration with RNA Clean & Concentrator™-5 (Zymo). We then made sequencing libraries from these samples as previously described [26,27]. We used the following antibodies: Bethyl A301-119A (Spen); Santa Cruz (G-14): sc-83849 (DKC1:V5); Abcam ab5642 (Ptbp1); Santa Cruz (3G6): sc-32315 (SAFA); Bethyl A300-674A (HnrnpK); customized LBR antibody from GenScript (LBR #4; 540774-1). DKC1 expressing plasmid has been deposited in GeneBank (accession number BankIt1965434 DKC1V5 KY070601).

Additional annotations

Cellular localization information (**Supplementary Table 2**) was retrieved from UniProt and LOCATE (*experimental* evidence; http://locate.imb.uq.edu.au/) databases. Expression levels in ES-E14 cell line were retrieved from ENCODE RNA-seq data averaging RPKMs of replicates with IDR<0.1.

1 Bellucci, M., Agostini, F., Masin, M. & Tartaglia, G. G. Predicting protein associations with long noncoding RNAs. *Nature methods* **8**, 444-445, doi:10.1038/nmeth.1611 (2011).

2 Agostini, F. *et al.* catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics* **29**, 2928-2930, doi:10.1093/bioinformatics/btt495 (2013).

3 Cirillo, D. *et al.* Neurodegenerative diseases: quantitative predictions of protein-RNA interactions. *RNA* **19**, 129-140, doi:10.1261/rna.034777.112 (2013).

4 Agostini, F., Cirillo, D., Bolognesi, B. & Tartaglia, G. G. X-inactivation: quantitative predictions of protein interactions in the Xist network. *Nucleic Acids Res* **41**, e31, doi:10.1093/nar/gks968 (2013).

5 Dassi, E. *et al.* AURA: Atlas of UTR Regulatory Activity. *Bioinformatics* **28**, 142-144, doi:10.1093/bioinformatics/btr608 (2012).

6 Siprashvili, Z. *et al.* Identification of proteins binding coding and non-coding human RNAs using protein microarrays. *BMC genomics* **13**, 633, doi:10.1186/1471-2164-13-633 (2012).

7 Kwon, S. C. *et al.* The RNA-binding protein repertoire of embryonic stem cells. *Nat Struct Mol Biol* **20**, 1122-1130, doi:10.1038/nsmb.2638 (2013).

8 Van Nostrand, E. L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature methods*, doi:10.1038/nmeth.3810 (2016).

9 McHugh, C. A. *et al.* The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature* **521**, 232-236, doi:10.1038/nature14443 (2015).

10 Chu, C. *et al.* Systematic discovery of xist RNA binding proteins. *Cell* **161**, 404-416, doi:10.1016/j.cell.2015.03.025 (2015).

11 Minajigi, A. *et al.* Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* **349**, doi:10.1126/science.aab2276 (2015).

12 Moindrot, B. *et al.* A Pooled shRNA Screen Identifies Rbm15, Spen, and Wtap as Factors Required for Xist RNA-Mediated Silencing. *Cell reports* **12**, 562-572, doi:10.1016/j.celrep.2015.06.053 (2015).

13 Monfort, A. *et al.* Identification of Spen as a Crucial Factor for Xist Function through Forward Genetic Screening in Haploid Embryonic Stem Cells. *Cell reports* **12**, 554-561, doi:10.1016/j.celrep.2015.06.067 (2015).

14 Grimm, S. The art and design of genetic screens: mammalian culture cells. *Nature reviews. Genetics* **5**, 179-189, doi:10.1038/nrg1291 (2004).

15 Lu, Q. *et al.* Computational prediction of associations between long non-coding RNAs and proteins. *BMC genomics* **14**, 651, doi:10.1186/1471-2164-14-651 (2013).

16 Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB* **6**, 26, doi:10.1186/1748-7188-6-26 (2011).

17 Muppirala, U. K., Honavar, V. G. & Dobbs, D. Predicting RNA-protein interactions using only sequence information. *BMC bioinformatics* **12**, 489, doi:10.1186/1471-2105-12-489 (2011).

18 Hasegawa, Y. *et al.* The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Developmental cell* **19**, 469-476, doi:10.1016/j.devcel.2010.08.006 (2010).

19 Cerase, A., Pintacuda, G., Tattermusch, A. & Avner, P. Xist localization and function: new insights from multiple levels. *Genome biology* **16**, 166, doi:10.1186/s13059-015-0733-y (2015).

20 Zhang, L. F., Huynh, K. D. & Lee, J. T. Perinucleolar targeting of the inactive X during S phase: evidence for a role in the maintenance of silencing. *Cell* **129**, 693-706, doi:10.1016/j.cell.2007.03.036 (2007).

21 Cerase, A. *et al.* Spatial separation of Xist RNA and polycomb proteins revealed by superresolution microscopy. *Proceedings of the National Academy of Sciences of the United States of America* **111**, 2235-2240, doi:10.1073/pnas.1312951111 (2014).

22 Sunwoo, H., Wu, J. Y. & Lee, J. T. The Xist RNA-PRC2 complex at 20-nm resolution reveals a low Xist stoichiometry and suggests a hit-and-run mechanism in mouse cells. *Proceedings of the National Academy of Sciences of the United States of America* **112**, E4216-4225, doi:10.1073/pnas.1503690112 (2015).
23 Takano, M. *et al.* The binding of lamin B receptor to chromatin is regulated by phosphorylation in the RS region. *European journal of biochemistry / FEBS* **269**, 943-953 (2002).
24 Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nature reviews. Genetics* **15**, 829-845, doi:10.1038/nrg3813 (2014).
25 Hall, L. L., Byron, M., Pageau, G. & Lawrence, J. B. AURKB-mediated effects on chromatin regulate binding versus release of XIST RNA to the inactive chromosome. *The Journal of cell biology* **186**, 491-507, doi:10.1083/jcb.200811143 (2009).
26 Engreitz, J. M. *et al.* The Xist lncRNA exploits three-dimensional genome architecture to spread across the X chromosome. *Science* **341**, 1237973, doi:10.1126/science.1237973 (2013).
27 Shishkin, A. A. *et al.* Simultaneous generation of many RNA-seq libraries in a single reaction. *Nature methods* **12**, 323-325, doi:10.1038/nmeth.3313 (2015).